



Stabilizing knowledge through standards a perspective for the humanities

Laurent Romary

► To cite this version:

Laurent Romary. Stabilizing knowledge through standards a perspective for the humanities. Going Digital - Evolutionary and Revolutionary Aspects of Digitization (Nobel Symposium - 147) - 2009, Jun 2009, Stockholm, Sweden. inria-00438724

HAL Id: inria-00438724

<https://inria.hal.science/inria-00438724>

Submitted on 4 Dec 2009

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Stabilizing knowledge through standards

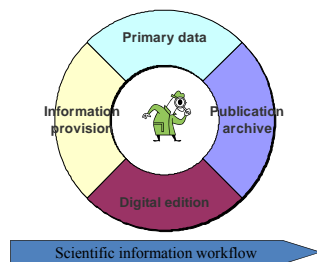
a perspective for the humanities

Laurent Romary
INRIA Gemo & Humboldt Univ. Berlin IDSL

Overview

- From scientific data to lexical databases
- Standardization — TEI, ISO, etc.
- Masculine, feminine, etc.
- Research infrastructures, libraries, etc.

The Scientist's (digital) ecology



04.12.2009

Seite 3

Working with research data

- Wide variety and complexity
 - High Energy Physics
 - Particle accelerators / colliders
 - Meteorology
 - Computer simulations
 - Astrophysics
 - Observations, stellar object descriptions
 - Biology
 - Spectrographic representations
 - Linguistics
 - Corpora, grammars, lexical databases

04.12.2009

Seite 4

“modern” dictionaries

Petit Larousse, 1905

Simple aims:

- Online rendering
- Precise queries on all fields
- Cross-reference with other dictionaries (dictionnaire de l'Académie)

(source: H. Manuélian, Métadif)



“old” dictionaries

Joachim Heinrich Campe
„Wörterbuch der deutschen Sprache“, 5
volumes, Braunschweig 1807–1811
„Wörterbuch zur Erklärung und
Verdeutschung der unserer Sprache
aufgedruckten fremden Ausdrücke.
Ein Ergänzungsband zu Adelungs und
Campe's Wörterbüchern“, Braunschweig
1813

Objective

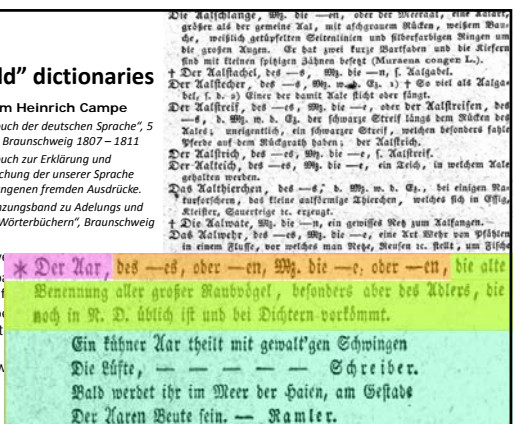
• 6000 p.

entries: 1

• Testb

Project

(Source W



Full-form lexica

Trésor de la Langue Française - Morpholou

- 539 413 inflected forms, 68 075 lemmas
- Natural Language Processing applications

(Source S. Alt, ATILF-CNRS)

```
chat sms,chat
...
chats smp,chat
...
cheik sms,cheikh:cheik
cheikh sms,cheikh:cheik
...
ferme axs sfs,ferme
ferme ipl1 ip3s spls sp3s im2s,fermer v
fermentent h,fermentent
ferment ip3p sp3p,fermer v
ferment sms,ferment
...
```

Multext-East lexicon

MSD	Feature structure	Table 6. Lexical MSDs (17)	Lexical types/examples	
Atf-p	Adjective Type-qualitative Degree-positive	41	3	бел, нигро, га
Atf-p-p	Adjective Type-qualitative Degree-positive Number-plural	5764	920	белы/белый, нигро/нигроый, га/гаый, эффекты/эффективный, эффекты/эффективный
Atf-pn	Adjective Type-qualitative Degree-positive Number-plural Case-nominative	3512	239	белые/белой, нигро/нигроый, га/гаый, эффекты/эффективный, эффекты/эффективный
Atf-pnf	Adjective Type-qualitative Degree-positive Number-plural Case-nominative	22089	4601	белые/белой, нигро/нигроый, га/гаый, эффекты/эффективный, эффекты/эффективный
Atf-pg	Adjective Type-qualitative Degree-positive Number-plural Case-genitive	5027	271	белых/белой, нигро/нигроый, га/гаый, эффекты/эффективный, эффекты/эффективный
Atf-pgf	Adjective Type-qualitative Degree-positive Number-plural Case-genitive	27346	4195	белых/белой, нигро/нигроый, га/гаый, эффекты/эффективный, эффекты/эффективный
Atf-pd	Adjective Type-qualitative Degree-positive Number-plural Case-dative	381	72	белым/белой, нигро/нигроый, га/гаый, эффекты/эффективный, эффекты/эффективный
Atf-pdf	Adjective Type-qualitative Degree-positive Number-plural Case-dative	2550	860	белым/белой, нигро/нигроый, га/гаый, эффекты/эффективный, эффекты/эффективный
Atf-pa	Adjective Type-qualitative Degree-positive Number-plural Case-accusative	1027	143	белых/белой, нигро/нигроый, га/гаый, эффекты/эффективный, эффекты/эффективный
Atf-paf	Adjective Type-qualitative Degree-positive Number-plural Case-accusative	9915	1919	белых/белой, нигро/нигроый, га/гаый, эффекты/эффективный, эффекты/эффективный
Atf-pl	Adjective Type-qualitative Degree-positive Number-plural Case-localive	735	117	белых/белой, нигро/нигроый, га/гаый, эффекты/эффективный, эффекты/эффективный
Atf-plf	Adjective Type-qualitative Degree-positive Number-plural Case-localive	3057	1488	белых/белой, нигро/нигроый, га/гаый, эффекты/эффективный, эффекты/эффективный

Jezikoslovno označevanje slovenščine <http://nl.ijs.si/jos>



Project JOS: Linguistic Annotation of Slovene

The JOS project is developing Slovene annotated corpora and associated resources meant to facilitate development in Human Language Technologies for the Slovene language. Current results include the JOS morphosyntactic specifications (tagset, definitions), two word-level annotated corpora, and two Web services. The developed resources are available under the Creative Commons license.

JOS annotated corpora

The JOS corpora contain sampled paragraphs from the *Friderik's* corpus, annotated with correct disambiguated morphosyntactic designations and lemmas. The *pos100k* corpus contains 100,000 words and has been extensively manually validated, while *pos1M* contains 1 million words with partially manually validated annotations.

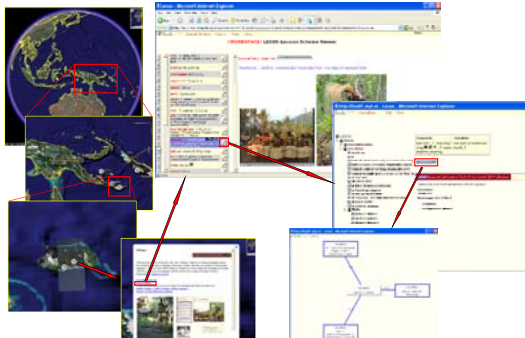
Both corpora are available in XML and decompiled tabular files and TEI header in HTML. The XML schemas are based on TEI P5 and the XML corpora contain the TEI header (metadata), both in Slovene and English, and samples from the texts. Text samples are individual paragraphs, each composed of sentences, and these in turn of words, punctuation and whitespace. Words are then annotated with their MSD and lemma. The XML corpora contain Slovene language MSDs and feature-structure lexicons, only, although the conversion to English language MSDs should be simple using the JOS MSD conversion tables.

The tabular files are smaller and probably easier for direct use, as they do not contain the header, and the token-level annotation is presented in tabular format. Each line in a file is thus a structural XML tag (e.g. <w> or <w> with TAB separated lemma and MSD annotations. These files do, however, lose the information on the original spacing, and, in the case of *pos1M* on whether the annotation is manual or automatic. The tabular files are available both with English and Slovene MSDs.

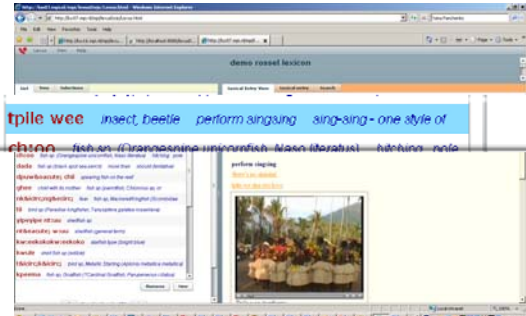
The corpora are available under the [Creative Commons Attribution-NonCommercial 4.0 license](http://creativecommons.org/licenses/by-nc-sa/4.0/), meaning that you are free to use it for any non-commercial purpose, provided that you give the original authors credit in scientific publications. This means citing the relevant publication or publications, referred to in the *background* part of this page.

- JOS project home
- Morphosyntactic specifications
- pos100k corpus
- pos1M corpus
- Web Services
- Publications
- Further links

Memory of endangered languages



Multimodal lexical information



Why standardizing all this?

- Defining methods or models to facilitate
 - Exchange of data
 - Pooling data from various origins
 - Interoperability between software components
 - Comparability of results
- Involves
 - From a scientific and technological point of view
 - Stabilizing/documenting existing practices, knowledge
 - Looking ahead for potential roadblocks (generalizations)
 - From an organizational point of view
 - International consensus, long term availability and maintenance

Standards: a complex picture

- Standardization bodies or consortia
 - National: AFNOR, ANSI, BSI, DIN, MSA, SIS (Swedish Standard Institute)
 - International: ISO, IEC, CEN, W3C, OASIS, TEI
- Specific fora
 - Many! e.g.
 - LISA (Localization Industry Standards Association)
- Projects with a pre-normative purpose
 - e.g. in Europe:
 - EAGLES, Multext, MATE, ISLE, Lirics, Kyoto

Can scientists bear standards?

- Standards are essentially “bad” for scientists
 - Freezing knowledge
 - Lost of time (which could be dedicated to research)
 - Forcing diverging views to agree
 - ...especially if the work is done by others
 - [also known as NIH syndrome: “not invented here”]
 - Forcing one to make data readable by others
 - ...

04.12.2009

Seite 14

How to answer reluctance?

- Main issues
 - Managing the trade-off between *interoperability* and *variability* of linguistic representation
 - Documenting and maintaining document formats
 - Unifying the management, query and presentation of linguistic resources
- A possible answer
 - Standards as specification platforms
- Major factors
 - Expressing constraints on models, adaptation to use cases
 - Identifying generic structures, preventing representation silos

Standardization for language resources: current state

- TEI
 - Initiated in 1987, driving force behind XML creation
 - P5 edition of the guidelines
 - Cf. specification platform (ODD)
- ISO
 - ISO/TC 37: Terminology and language resources
 - ISO/TC 37/SC 2: ISO 639 series (language codes)
 - ISO/TC 37/SC 3: ISO 16642 (Terminology)
 - ISO/TC 37/SC 4: Language resource management (2002)
- W3C
 - ITS (Internationalization (I18n) activity)
 - SMIL Text (Synchronized Multimedia Integration Language)

Intermezzo — an XML tutorial

- XML is about awful angle brackets

```
<gramGrp>
  <gen>f</gen>
  <num>p</num>
</gramGrp>
```

- XML is about trees



- Issues

- Specifying structures
- Providing semantics

Modeling Lexical Structures with the TEI

How it all started



TEI example

```
<stage>Enter Barnardo and Francisco, two Sentinels, at several doors</stage>
<sp who="Barnardo">
  <l part="f">Who's there?</l>
</sp>
<sp who="Francisco">
  <b>Nay, answer me. Stand and unfold yourself.</b>
</sp>
<sp who="Barnardo">
  <l part="f">Long live the king!</l>
</sp>
<sp who="Francisco">
  <l part="m">Barnardo?</l>
</sp>
<sp who="Barnardo">
  <l part="f">He. </l>
</sp>
```

04.12.2009

Seite 20

Following the TEI spirit

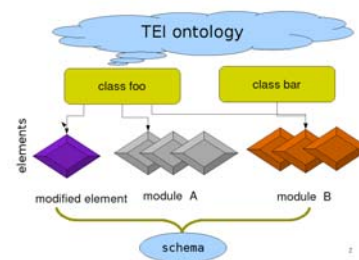
Conformance to the TEI means:

- Sharing a common text encoding culture
- Sharing the same vocabulary (when applicable)
- Allowing user autonomy in defining modifications (extensions, customization), but sharing the mechanisms to do so

04.12.2009

Seite 21

TEI architecture — playing Lego



04.12.2009

Seite 22

List of TEI Modules		
	Module name	A short description
add	analysis	Simple analytic mechanisms
add	certainty	Certainty and uncertainty
add	core	Elements common to all TEI documents
add	corpus	Corpus texts
add	dictionaries	Printed dictionaries
add	drama	Performance texts
add	figures	Tables, formulae, and figures
add	gali	Character and glyph documentation
add	header	The TEI Header
add	iso-fs	Feature structures
add	linking	Linking, segmentation and alignment
add	msdescription	Manuscript Description
add	namesdates	Names and dates
add	nets	Graphs, networks, and trees
add	spoken	Transcribed Speech
add	tagdocs	Documentation of TEI modules
add	textfmt	Critical Apparatus
add	textstructure	Default text structure
add	transcr	Transcription of primary sources
add	verse	Verse structures

Encoding a dictionary entry

```
<entry>
  <form>
    <orth>table</orth>
  </form>
  <def>Pièce de mobilier.
  <cit>
    <quote>Une table de cuisine</quote>
  </cit>
</entry>
```

Annotations in the diagram:

- Selecting content:** Points to the `<orth>` element.
- Constraining content:** Points to the `<pos>`, `<gen>`, `<num>`, and `<tense>` elements.
- Adding content:** Points to the `<def>` element, with an example: `e.g.: <transitivity>`.

Inflectional variants

* Der Kar, des —es, oder —en, Mz. die —e, oder —en, die alte Benennung aller großer Raubvögel, besonders aber des Adlers, die noch in N. D. üblich ist und bei Dichtern vorkommt.

Der Aar, des -es, oder -en,	<pre> <form type="inflected"> <gramGrp> <case>genitive</case> <number>singular</number> </gramGrp> <form type="determiner"> <orth>des</orth> </form> <form type="headword"> <orth> <ovar><ovar/>-es</ovar> </orth> </form> ... </form> </pre>
---	--

Folie 25

Specification and documentation

TEI's literate programming with ODD (One Document Does it all) provides: schema specification (DTD, RelaxNG, W3C), user oriented documentation, modularity, classes, extensibility.

<gen> (gender) identifies the morphological gender of a lexical item, as given in the dictionary. [9.3.1 Information on Written and Spoken Forms](#)

[illegible]

Before we go any further...

- Which normative reference for the values of element like <gen> (grammatical gender)?
 - Not an issue specific to dictionary design
 - Cf. linguistic annotation at large (e.g. POS tagging)
 - Not an issue specific to the TEI community
 - Such values and their semantics should be defined independantly of any specific tagset
- Is <gen> a self-standing notion?

Modeling Lexical Resources within ISO/TC 37/SC 4

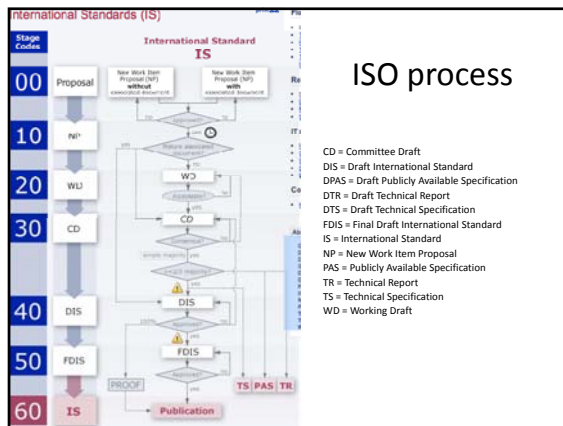
ISO in short



- International Organization for Standardization (<http://www.iso.org>)
 - Administrative view
 - Federation of national standardization bodies
 - Technical view
 - Organized in technical committee and sub-committees
 - ISO technical committees

ISO: a standardisation body

- Providing unique references
 - Language (ISO 639), country (ISO 3166) and script coding (ISO 15924)
 - zh-SG (Chinese for Singapore)
 - sr-Cyrl (Serbian written with Cyrillic script)
- Providing definitions and principles
 - Character encoding
 - ISO 636, ISO 8859-x, ISO 10646/Unicode
- Standard as an evolving material



ISO/TC 37/SC 4 projects

ISO 24610-1:2006 Feature structures -- Part 1: Feature structure representation	ISO/CD 24612 Linguistic annotation framework (LAF)
ISO/DIS 24610-2 Feature structures -- Part 2: Feature system declaration	ISO/NP 24619 Citation of Electronic Resources (CERER)
ISO 24613:2008 Lexical markup framework (LMF)	ISO/WD 24616 Multi lingual information framework (MLIF)
	ISO/DIS 24611 Morpho syntactic annotation framework (MAF)
	ISO/CD 24615 Syntactic annotation framework (SynAF)
ISO/DIS 24617-1 Semantic annotation framework (SemAF) -- Part 1: Time and events	ISO/CD 24614-1 Word segmentation of written texts for mono-lingual and multi-lingual information processing -- Part 1: General principles and methods
ISO/CD 24617-2 Semantic annotation framework (SemAF) -- Part 2: Dialogue acts	ISO/WD 24614-2 Word segmentation of written texts for mono-lingual and multi-lingual information processing -- Part 2: Word segmentation for Chinese, Japanese and Korean

General modeling framework

- Meta-model
 - General, underlying model that informs current practice
- Data-categories
 - Provides the elementary descriptors to instantiate models

Application to lexical structures

LMF — Lexical Markup Framework
(ISO 24613)

LMF as an ISO project

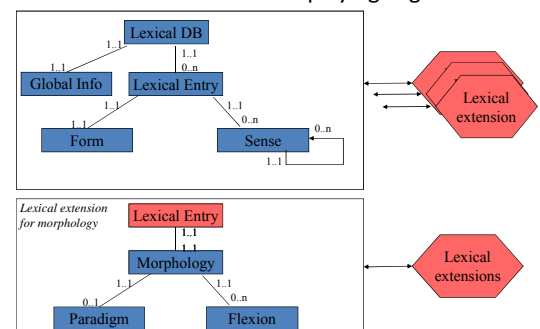
- Summer 2003: new work item proposal (US) delegation
- Fall 2003: technical proposal (FR) for a data model dedicated to NLP lexica
- ISO 24613
 - Convenor:
 - Nicoletta Calzolari (IT)
 - Editors:
 - Gil Francopoulo (FR), Monte George (US)
 - 13 versions written, dispatched (to the National delegations nominated experts), commented and discussed in various ISO technical meetings
- IS (= published standard) in oct. 2008

Tubingen 2007

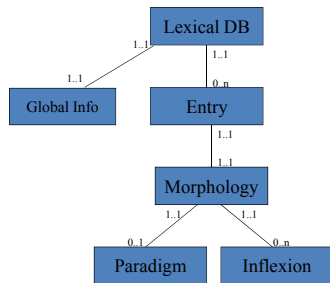
Lex-Sem & Onto-Resources

35

LMF architecture — playing Lego

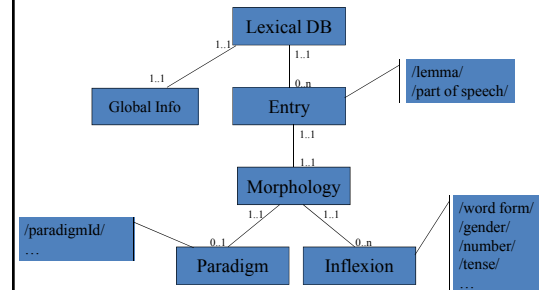


Example: designing a full-form lexicon



Seite 37

Decorating the model



Seite 38

A possible XML instance

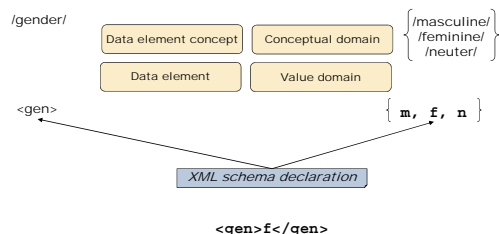
```
<lexicalEntry>
  <lemma>chat</lemma>
  <grammaticalCategory>noun</grammaticalCategory>
  <morphology>
    <paradigm>
      <paradigmIdentifier>fr-s-plural</paradigmIdentifier>
    </paradigm>
    <inflexion>
      <wordForm>chat</wordForm>
      <number>singular</number>
    </inflexion>
    <inflexion>
      <wordForm>chats</wordForm>
      <number>plural</number>
    </inflexion>
  </morphology>
</lexicalEntry>
```

Seite 39

A central concept: data category

- Definition
 - Elementary descriptor used in a linguistic description or annotation scheme
- Examples
 - Fields: /part of speech/, /grammatical gender/
 - Values: /feminine/, /plural/, /dual/, /ablative case/
- Role
 - Specification
 - Documentation
- A reference space for schema designers
 - Towards an international registry for language resources
 - Data Category Registry (DCR); cf. ISO 12620

Formal background: ISO 11179



Some deeper thoughts on *gender*

- A central category in linguistic and computational linguistic
 - Lexica, morpho-syntactic tagging, agreement in syntax, etc.
- Can we standardize “gender”
 - Interoperability vs. language variety
- By the way, gender is not exactly “sex”
 - ISO 5218, Information technology — Codes for the representation of human sexes
 - 0 = not known; 1 = male; 2 = female; 9 = not applicable

The linguistic view

- What is gender:
 - “a classification of nominals, as shown by agreement”
 - E.g. die Katze – der Hund
 - Determiners, adjectives, numerals, verbs
 - E.g. Control by anaphoric pronouns (cf. en)
 - Die Katze... sie...
 - Not present in all languages
- [Number of genders](#) (Greville G. Corbett)

Application: Independent personal pronouns

- Example: Rif Berber (McClelland 2000: 27)

1sg	naš	1pl	našnin
2sg.m	š a k	2pl.m	k a niw
2sg.f	šam	2pl.f	kanint
3sg.m	natta	3pl.m	nitnin
3sg.f	nattaθ	3pl.f	nitanti

- [Gender Distinctions in Independent Personal Pronouns](#), Source: Anna Siewierska (cf. wals.info)

The TC 37 model — ISO 12620

Entry Identifier:	grammatical gender		
Profile:	morpho-syntax		
Definition (fr):	Catégorie grammaticale reposant, selon les langues et les systèmes, sur la distinction naturelle entre les sexes ou des critères formels (Source: TLFi)		
Object Language: fr	Object Language: en	Object Language: de	
Name: genre	Name: gender	Name: Geschlecht	
Conceptual Domain: {feminine/, /masculine/}	Name: grammatical gender	Name: Genus	Conceptual Domain: {feminine/, /masculine/, /neuter/}

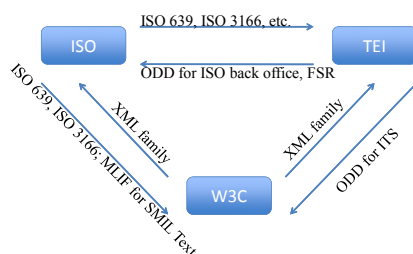
Convergence?

Petit Larousse 1905 by Métadiff (source H. Manuélian) → goes TEI

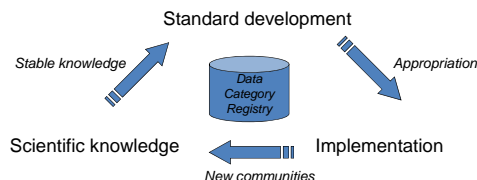
```

<entry>
  <form> Campe by Uni. Würzburg (W. Wegstein) → goes TEI
  <ortho>
  <prom>
  </form>
  <sense>
  <def t> Morphalou 2005 by ATILF (S. Alt) → goes LMF
  </sense>
  <gram>
  <pos>
  <gran>
  <def>
  </def>
  </gran>
  </pos>
  </sense>
  </gram>
  </entry>
  <lemma> cheikh</lemma>
  <spellingVariant> cheik</spellingVariant>
  <cit t>
  <qu>
  </qu>
  <grammaticalCategory> common noun</grammaticalCategory>
  <grammaticalGender> masculine</grammaticalGender>
  <bil>
  </bil>
  <wordForm> cheikh</wordForm>
  <cit i>
  </cit i>
  <wordForm> cheikhs</wordForm>
  <grammaticalNumber> plural</grammaticalNumber>
  </wordForm>
  </entry>
  </morphology>
  </lexicalEntry>
  
```

Convergence?



Standards as an emanation from scientific knowledge



Epilogue

RESEARCH INFRASTRUCTURES IN THE HUMANITIES

Research Infrastructures

- In general: [permanent](#) and [physical](#)
- Natural sciences: ice breakers for polar research, satellites, telescopes, particle accelerators, laboratories
- RIs for the humanities?
 - **Cultural heritage** in all forms is the main source of humanities research
 - **Libraries and archives** are the traditional “laboratories” for the humanities
- In the digital age, essential for innovative humanities research is:
 - [Access](#) to digitised heritage data (data bases, text corpora, speech, image collections, etc.)
 - [Tools](#) to process this information

04.12.2009

Seite 50

Core activities

- Digitise – Curate – Preserve
 - Standards development and promotion
 - Curation, preservation and digitisation services
 - Technology platforms
 - Legal services and advice
- Discover – Access – Deliver
 - Authentication and authorisation,
 - Harvesting, aggregating, hosting
 - User-friendly discovery, delivery and use
- Connect – Collaborate – Use
 - Supporting communities of practice
 - Facilitating new research practice
 - Tools and registries

04.12.2009

Seite 51

(conclusive) priorities

- Mastering the technology
 - Not all scientist are technological geeks
 - Transparency
- Answering priority needs
 - Strong request to provide infrastructures for simple types of data
 - Pragmatic sense
- Preserving scientific patrimony
 - High amounts of research data is continuously lost
 - Identification, preservation

04.12.2009

Seite 52

Should we/you be afraid of standards?

```
<cit>
  <quote>Yes you should be afraid, but you should be
    more afraid of not having them</quote>
  <author>Wendell Piez</author>
</cit>
```